**RII - Tarefa 2**

**Extração de Informação Textual**

**Descrição geral:**

Trabalho em duplas

**Tarefa:** realizar Extração de Informação a partir dos documentos selecionados com auxílio do professor. Construir um Wrapper para extrair informação fixa, baseada em tags

* + Construído por vocês (podem usar aprendizagem de máquina ou regras).
  + Testar com 10 documentos - Eu vou fornecer os documentos.

**Exemplo de dados para extração com wrapper**

Exemplo: Ver sites de classificados

<https://www.jornalclassificados.com.br/classificados/veiculos>

### Anúncio:

<https://www.jornalclassificados.com.br/classificados/veiculos?it=idea-14-2012&id=262289>

## IDEA ,14 2012  anúncio patrocinado

#### Carros

#### R$26.500,00

Motor Fire atractive prata completo com trava carneiro de segurança aceito troca, particular

### Formulário extraído:

* Marca: Fiat (campo ausente, mas pode ser obtido através de consulta na Web)
* Modelo: IDEA Fire attractive
* Preço: 26500,00
* Motor: 1.4 (provavelmente, já que apareceu “14” junto ao nome do carro)
* Ano: 2012
* Km:
* Tipo de Combustível:
* Câmbio:
* Direção:
* Cor: prata
* Ar-cond: sim
* N-portas:
* Opcionais: trava, carneiro de segurança
* Outros: aceito troca, particular
* Telefone para contato: Fixo: (0)  Cel:(97) 4125-435

## FOX SUNRISE - 2010 - 1.0

#### Carros

#### R$21.900,00

FOX SUNRISE 2010 - 1.0 4 PORTAS\*COMPLETO\*BRANCO\*ÚNICO DONO\* IMPECÁVEL\*KM 82000 WHATSAPP 19 9 97159968 Obrigado, confirmar recebimento

### Formulário extraído:

* Marca: VW (campo ausente, mas pode ser obtido através de consulta na Web)
* Modelo: Fox Sunrise
* Preço: 21900,00
* Motor: 1.0
* Ano: 2010
* Km: 82000
* Tipo de Combustível:
* Câmbio:
* Direção:
* Cor: Branco
* Ar-cond: sim (note que o anuncio diz ‘completo’)
* N-portas: 4
* Opcionais: completo
* Outros: único dono, impecável
* Telefone para contato: Fixo: (0)  Cel:(97) 4125-435

**Projeto & Relatório** (tudo misturado):

Escrevam um relatório simples, com as informações solicitadas abaixo:

# Introdução

# Nome dos membros da equipe e breve descrição da tarefa realizada.

# Descrição dos documentos (corpus) de entrada, com 2 exemplos

# Template de saída (completo para equipes de 3 alunos para automóveis)

         Marca: (WV, Ford, GM, Fiat, etc...)

         Modelo: (Fox, Fiesta, Uno, etc)

         Preço: valor numérico

         Motor: 1.0, 1.3, 1.6, ...

         Ano: cadeia numérica indicativa de ano

o   Se o ano não vier informado na 1ª linha, vejam se está no campo “Descrição”.

o   Se o ano vier com 2 dígitos, completem para quatro dígitos.

o   Se o ano vier no formato “98/99”, podem extrair desse jeito mesmo.

         Km: cadeia numérica

         Cor: cadeia alfa indicativa de cor

o   Lembrem-se de que a cor pode vir com 2 nomes “azul metálico”, “azul claro”

         Combustível: (Flex, gasolina, álcool, diesel, GVN)

         Câmbio: (Manual, automático)

         Direção: (Mecânica, hidráulica)

         Ar-cond: (sim, não)

         N-portas

         Opcionais:

o   Opcionais do automóvel: banco de couro, radio, CD e MP3 Player, alarme, airbag, ...

         Outros:

o   Exemplo: único dono, pouco rodado...

         Telefones: fixo e celular

Obs. 1: os campos que não forem extraídos porque (1) não estão informados no anúncio, ou (2) existem, mas o sistema não conseguiu extrair, devem ficar em branco.

Obs. 2: O template será mais simples para quem fizer o trabalho em equipes de 2 membros ou sozinho. Os campos em azul (tipo de combustível, câmbio, direção e ar-condicionado) não precisam ser extraídos/preenchidos.

# Arquitetura do sistema

Apresentar uma figura simples com os módulos do sistema. Prover uma descrição breve de cada módulo, e informar qual a técnica usada na construção do wrapper.

# Testes e avaliação dos resultados

A avaliação dos resultados terá que ser feita de forma manual. Neste caso, não temos matriz de relevância montada a priori ☺

Vamos medir P e C para CADA CAMPO do template. Isso é mais informativo do que calcular P e C para o template como um todo.

* **Precisão:** total de vezes que o campo x foi preenchido corretamente dividido pelo total de vezes que o campo foi preenchido pelo seu sistema. Queremos ver se o sistema acerta na extração.
* **Cobertura:** total de vezes que o campo x foi preenchido corretamente dividido pelo total de vezes que o campo foi informado nos anúncios. Queremos ver se o sistema deixa de extrair campos informados.

Incluir no relatório uma matriz de resultados para CADA módulo do sistema.

**Exemplo de Matriz de resultados para o Wrapper**

Exemplo para o campo 1:

* Considere uma base de documentos com **10 docs** de entrada
* O campo 1 está presente em 7 documentos
* O sistema extraiu o campo 1 corretamente apenas 4 vezes
* O sistema preencheu o campo 1 erroneamente 1 vez
* O sistema não conseguiu extrair o campo 3 vezes

|  |  |  |
| --- | --- | --- |
|  | Precisão | Cobertura |
| Campo 1 | **4/5** – o campo 1 foi preenchido 5 vezes, porém o sistema só acertou 4 vezes. | **4/7** – o campo foi preenchido 4 vezes, porém ele estava presente em 7 documentos de entrada |
| Campo 2 |  |  |
| Campo 3 | ... | ... |
| Campo 4 |  |  |
| Campo 5 |  |  |
| Média |  |  |

# Conclusão

Um texto curto para explicar o que vocês concluem a partir do resultado dos experimentos.